

# Appunti di Econometria

## ARGOMENTO [5]: ANALISI DEI DATI PANEL

Maria Luisa Mancusi – Università Bocconi

Novembre 2009

### 1 I dati panel

Un panel è un campione che contiene osservazioni su  $N$  individui per  $T$  anni. Le osservazioni su ogni individuo sono, cioè, ripetute nel tempo ( $\rightarrow$  serie storica su ogni individuo). Vediamo qualche esempio.

- Dati d'impresa:

- funzione di produzione

Es. Si hanno i seguenti dati su 576 imprese del settore manifatturiero per il periodo 1985-1994:

$s$  = fatturato in milioni di euro ( $\rightarrow$  output)

$e$  = numero di lavoratori in 1000 ( $\rightarrow$  lavoro)

$k$  = impianti e macchinari in milioni di euro ( $\rightarrow$  capitale)

- investimento e dati finanziari

- innovazione (brevetti, R&S, ...)

- Dati su  $N$  famiglie per  $T$  anni:

- scelte di consumo (es. *Panel Survey on Income Dynamics*, PSID)

- Dati macroeconomici su  $N$  paesi per  $T$  anni:

- PIL, occupazione, export, ecc.

In particolare, le caratteristiche dei panel “microeconomici” sono:

- il numero di individui (la dimensione cross-section) è tipicamente grande
- la serie storica per ogni individuo è solitamente breve

Il beneficio principale dei dati panel è che essi consentono di rispondere a domande a cui non si può invece dare risposta quando si utilizza un campione cross-section o una serie storica.

$\Rightarrow$  Es. 1: *cross-section vs. dati panel*

In un dato anno osserviamo che il 30 per cento di un campione di imprese realizza una o più innovazioni. Due sono le interpretazioni possibili:

- a) ogni anno, in media il 30 per cento di imprese realizza una o più innovazioni
- b) le stesse imprese, che rappresentano il 30 per cento del campione, innovano ogni anno.

⇒ *Es. 2: serie storica vs. dati panel*

Da un'analisi della serie storica degli investimenti in R&S delle imprese risulta che il loro tasso di crescita annuo è pari al 2 per cento. Questo potrebbe essere il risultato di una crescita annua del 2 per cento in tutte le imprese o, ad esempio, di una crescita annua del 4 per cento in circa la metà delle imprese e di una crescita nulla nella restante metà.

In breve, la disponibilità di dati panel ci permette di tenere conto dell'eterogeneità degli individui.

Dunque, il campione di cui si dispone è:

$$(y_{it}, x_{it}) \quad \begin{array}{l} i = 1, \dots, N \\ t = 1, \dots, T_i \end{array}$$

$$\begin{array}{ll} T_i = T & \forall i \Rightarrow \text{PANEL BILANCIATO} \\ T_i \neq T_j \text{ per qualche } i \neq j & \Rightarrow \text{PANEL NON BILANCIATO} \end{array}$$

Noi supporremo sempre che il panel sia bilanciato. I metodi di stima che studieremo (e le formule associate) possono essere facilmente adeguati all'analisi di panel non bilanciati se la causa per cui la dimensione temporale è diversa per diversi individui è di tipo esogeno. Se, al contrario, la causa è di tipo endogeno, sono spesso necessari metodi di stima più complessi. Ad esempio, negli ultimi trent'anni negli Stati Uniti ad un campione di circa 10000 individui è stato ripetutamente (ogni anno) sottoposto un questionario con lo scopo di raccogliere dati sull'evoluzione dei redditi e dei consumi delle famiglie. Il risultato è un'enorme dataset di tipo panel noto come Panel Survey on Income Dynamics. Mentre per alcuni individui sono disponibili dati per l'intero periodo (dall'anno in cui il questionario è stato sottoposto per la prima volta all'anno corrente), per altri la serie storica dei dati risulta più breve. Ciò può essere dovuto a cause di tipo puramente esogeno (es. alcuni di questi individui sono deceduti), ma anche a cause di tipo endogeno, cioè strettamente collegate ai fenomeni ed alle variabili studiate. Ad esempio, supponiamo che per rispondere al questionario sia necessaria un'ora e che, ogni anno, agli individui è corrisposta una somma pari a 50\$ per rispondere al questionario e partecipare così all'indagine. Se nel corso del tempo uno di questi individui è diventato un avvocato e guadagna 250\$ all'ora, può decidere di cestinare il questionario e non partecipare più all'indagine. La serie storica disponibile per questo individuo sarà più breve per motivi legati all'oggetto dell'indagine e dell'analisi. In breve, se gli individui più ricchi sono anche quelli che, con maggiore probabilità, abbandonano l'indagine, il campione non sarà più "casuale" e può essere necessario apportare delle correzioni perché le stime basate su tale campione risultino non distorte. Dunque, anche se noi non ci occuperemo di questo problema, è bene tenere presente che, se si ha un panel non bilanciato, la prima cosa da controllare è la causa per cui è tale.

## 2 Il modello lineare statico con dati panel

Il modello lineare con dati panel è specificato in modo analogo al modello lineare su dati cross-section con la sola differenza che ora dovremo tener conto del fatto che la variabilità è sia tra individui che nel tempo:

$$y_{it} = x'_{it}\beta + \varepsilon_{it} \quad i = 1, \dots, N; t = 1, \dots, T$$

Convenzionalmente, le osservazioni sono ordinate nel seguente modo:

$$y = \begin{pmatrix} y_{11} \\ y_{12} \\ \dots \\ y_{1T} \\ y_{21} \\ \dots \\ y_{2T} \\ \dots \\ y_{N1} \\ \dots \\ y_{NT} \end{pmatrix} \rightarrow \left\{ \begin{array}{l} i = 1 \\ \\ \\ i = 2 \\ \\ \\ \\ \\ i = N \end{array} \right.$$

Dunque, in forma matriciale il modello lineare è:

$$y_{NT \times 1} = X_{NT \times K} \beta_{K \times 1} + \varepsilon_{NT \times 1}$$

Per tenere conto dell'eterogeneità degli individui, ossia di caratteristiche peculiari di ciascun individuo che, presumibilmente, non siamo in grado di osservare e che permangono nel tempo, l'errore viene specificato nel seguente modo:

$$\varepsilon_{it} = \alpha_i + u_{it}$$

$\alpha_i$  : effetto individuale, costante nel tempo (eterogeneità persistente non osservata)

$u_{it}$  : errore i.i.d. su  $i$  and  $t$  (cioè tra gli individui e nel tempo)

Esempio:

Siamo interessati a stimare una funzione di produzione di tipo Cobb-Douglas utilizzando i dati di impresa citati nella sezione precedente. Il modello da stimare è:

$$\log s_{it} = \beta_1 \log k_{it} + \beta_2 \log e_{it} + \varepsilon_{it} \quad i = 1, \dots, 576; \quad t = 1985, \dots, 1994$$

dove

$$\varepsilon_{it} = \alpha_i + u_{it}$$

$$y_{it} = \log s_{it} \text{ ed}$$

$$x_{it} = \{\log k_{it}, \log e_{it}\}$$

$\alpha_i$  : differenze persistenti nella produttività delle imprese (es. dovute ad abilità manageriali, potere di mercato, ecc.)

$u_{it}$  : differenze transitorie nella produttività delle imprese (es. dovute a shock di domanda/offerta nella località in cui opera l'impresa, ecc.)

I dati sono organizzati nel modo seguente:

i	t	logs	loge	logk
1	1985	4,827,377	0,547543	3,636,058
1	1986	4,967,567	0,547543	3,744,219
1	1987	5,117,173	0,585562	3,942,862
1	1988	5,192,134	0,518794	4,081,715
1	1989	5,280,092	0,639746	4,249,138
1	1990	5,559,919	0,747162	4,397,481
1	1991	5,683,032	0,887068	4,553,508
1	1992	5,757,310	0,834213	4,760,283
1	1993	5,902,956	0,900974	4,847,653
1	1994	6,106,432	0,972293	4,987,216
2	1985	4,653,093	0,368109	3,344,380
2	1986	4,685,662	0,442761	3,592,258
...	...	...	...	...

Il metodo di stima più appropriato dipende dalle ipotesi che facciamo su  $\alpha_i$ .

## 2.1 Effetti individuali fissi

$$y_{it} = x'_{it}\beta + (\alpha_i + u_{it}) \quad i = 1, \dots, N; \quad t = 1, \dots, T$$

$$\alpha_i \text{ fissa}, \forall i$$

$$y_{it} = \alpha_i + x'_{it}\beta + u_{it}$$

$$u_{it} \text{ i.i.d. } N(0, \sigma_u^2) \quad (N.B. \text{ su } i \text{ e } t)$$

Assumere che  $\alpha_i$  è fissa equivale ad ipotizzare che le differenze tra individui sono catturate da differenze nella costante. Il modello iniziale è dunque equivalente ad un modello in cui le variabili esplicative sono le  $x$  ed  $N$  dummies, una per ogni individuo<sup>1</sup> :

$$y_{it} = \sum_{j=1}^N \alpha_j d_{ij} + x'_{it}\beta + u_{it}$$

$$d_{ij} = \begin{cases} 1 & \text{se } j = i \\ 0 & \dots \text{se } j \neq i \end{cases}$$

In forma matriciale:  $y = D\alpha + X\beta + u$ , dove  $D$  è la matrice  $NT \times N$  contenente le dummies.

L'insieme di parametri da stimare è dunque:  $(\beta, \alpha_1, \dots, \alpha_N, \sigma_u^2)$  e la stima può essere ottenuta con OLS. Lo stimatore di  $\beta$  che ne risulta è detto *least squares dummy variable estimator* (LSDV). Questo metodo di stima diventa però impraticabile se il panel contiene un numero di individui molto elevato (es. in PSID  $N \simeq 10000!$ ). Fortunatamente, si può dimostrare che lo stesso stimatore di  $\beta$  si può ottenere stimando il modello in deviazioni dalle medie individuali, questo stimatore è detto *fixed effects estimator* (FE).

A partire dal modello  $y = D\alpha + X\beta + u$  si effettuano le seguenti regressioni:

<sup>1</sup>Supponiamo che le  $x$  non includano la costante, per evitare il problema della multicollinearità.

1. regressione di  $y$  su  $D$ , i cui residui sono  $y^* = y - D(D'D)^{-1}D'y = \left[ I - D(D'D)^{-1}D' \right] y = My$
2. regressione di  $X$  su  $D$  i cui residui sono  $X^* = MX$

Si ricordi ora che  $MD = 0$ , dunque  $My = MD\alpha + MX\beta + Mu$ , ovvero  $y^* = X^*\beta + u^*$ . Lo stimatore di  $\beta$  si ottiene quindi da una regressione OLS di  $y^*$  su  $X^*$ . Per capire in cosa consiste questo modello è sufficiente ricordare che:

$$M_{NT \times NT} = \begin{bmatrix} \overline{M} & 0 \\ 0 & \overline{M} \end{bmatrix} \text{ e } \overline{M}_{[TxT]} = I_T - \frac{1}{T}ii'$$

ovvero si stima con OLS il modello in deviazione dalle medie individuali<sup>2</sup>:

$$(y_{it} - \overline{y}_i) = (x_{it} - \overline{x}_i)' \beta + (u_{it} - \overline{u}_i)$$

dove:

$$\overline{y}_i = T^{-1} \sum_{t=1}^T y_{it}; \quad \overline{x}_i = T^{-1} \sum_{t=1}^T x_{it}; \quad \overline{u}_i = T^{-1} \sum_{t=1}^T u_{it}; \quad \overline{\alpha}_i = \alpha_i$$

Questo è ottenuto sottraendo dal modello originario il modello seguente, ottenuto dal primo facendone la media individuale:

$$\overline{y}_i = \alpha_i + \overline{x}_i' \beta + \overline{u}_i$$

In pratica si adotta una trasformazione che elimina le  $\alpha_i$  e si ottiene così lo stimatore per  $\beta$ :

$$\widehat{\beta}_{FE} = \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \overline{x}_i)(x_{it} - \overline{x}_i)' \right)^{-1} \left( \sum_{i=1}^N \sum_{t=1}^T (x_{it} - \overline{x}_i)(y_{it} - \overline{y}_i)' \right)$$

Successivamente si può impiegare la stima di  $\beta$  così ottenuta per stimare  $\alpha_i$  come residuo medio:

$$\widehat{\alpha}_{i,FE} = \overline{y}_i - \overline{x}_i' \widehat{\beta}_{FE}$$

La varianza di  $\widehat{\beta}_{FE}$  è  $Var(\widehat{\beta}_{FE}) = \sigma_u^2 (X'MX)^{-1}$  ed è stimata sostituendo a  $\sigma_u^2$  il suo stimatore consistente:

$$s_{FE}^2 = \frac{e'e}{NT - N - K} = \frac{1}{NT - N - K} \sum_{i=1}^N \sum_{t=1}^T (y_{it} - \widehat{\alpha}_{i,FE} - x_{it}' \widehat{\beta}_{FE})^2$$

Si può dimostrare che:

- $\widehat{\beta}_{FE}$  e  $\widehat{\alpha}_{i,FE}$  sono non distorti se  $E(u_{is}|x_{it}) = 0, \forall i, s, t$

<sup>2</sup>Se  $T = 2$  si ottiene semplicemente il modello in differenze:  $(y_{i2} - y_{i1}) = ecc.$

- $\widehat{\beta}_{FE}$  e  $\widehat{\alpha}_{i,FE}$  sono consistenti se  $E(x_{it}u_{is}) = 0, \forall i, s, t^3$

Mentre però  $\widehat{\beta}_{FE}$  è consistente per  $NT \rightarrow \infty$ , con, indifferentemente  $N \rightarrow \infty$  o  $T \rightarrow \infty$ ,  $\widehat{\alpha}_{i,FE}$  è consistente solo per  $T \rightarrow \infty$  (se  $N \rightarrow \infty$ , ma  $T$  rimane fisso, aumenta il numero di  $\alpha_i$  da stimare).

La distribuzione asintotica dello stimatore FE è normale, dunque le solite procedure di inferenza statistica possono essere utilizzate.

Se nel modello originario è inclusa una costante tra le variabili esplicative, ovvero il modello è:

$$y_{it} = \gamma + \alpha_i + x'_{it}\beta + u_{it}$$

per evitare il problema della perfetta multicollinearità originato dalla presenza di N variabili dummy tra loro complementari, occorre introdurre la seguente normalizzazione:

$$\sum_{i=1}^N \alpha_i = 0$$

Il modello viene quindi stimato nel modo seguente:

**Stadio 1:** Si ottiene  $\widehat{\beta}_{FE}$  stimando con OLS il modello trasformato (da cui viene eliminata anche la costante)

**Stadio 2:**  $\widehat{\gamma} = \bar{y} - \bar{x}'\widehat{\beta}_{FE}$ , dove  $\bar{y} = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T y_{it}$  e  $\bar{x} = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T x_{it}$

**Stadio 3:**  $\widehat{\alpha}_{i,FE} = \bar{y}_i - \bar{x}'_i \widehat{\beta}_{FE} - \widehat{\gamma}$

E' opportuno notare che in questo caso le  $\alpha_i$  hanno una diversa interpretazione: rappresentano le deviazioni dell'effetto individuale dalla media comune,  $\gamma$ .

Si noti che se gli effetti individuali, oltre ad essere fissi, sono in realtà uguali per tutti gli individui (cioè sono un effetto "comune"  $\gamma \leftrightarrow \alpha_i = 0$  ogni  $i = 1, \dots, N$ ), allora il modello originario può essere direttamente stimato con OLS. Lo stimatore di  $\alpha$  e di  $\beta$  che ne risulta è consistente ed efficiente. Questo approccio non distingue tra due individui diversi e lo stesso individuo in due istanti del tempo diversi  $\Rightarrow$  non è più accurato se esistono differenze tra gli individui ( $\alpha_i \neq \alpha_j$  per  $i \neq j$ ). Lo stimatore OLS risulta, infatti, distorto, perché si stanno omettendo delle variabili rilevanti (le dummies, appunto). La Figura 1 alla fine di queste dispense mostra la distorsione dello stimatore OLS nel caso di un'unica variabile esplicative. Questo evidenzia l'importanza di considerare esplicitamente l'eterogeneità degli individui (e quindi i benefici dei dati panel) al fine di avere una stima consistente dei parametri di interesse,  $\beta$ . Non è tanto l'interesse specifico nella stima di  $\alpha$  che giustifica la sua introduzione nel modello, quanto piuttosto le possibili conseguenze sulla stima di  $\beta$  derivanti dalla sua esclusione.

Se assumiamo che l'errore sia normalmente distribuito [ $u_i$  iid  $N(0, \sigma^2)$ ] è possibile testare la presenza di effetti individuali diversi nel modo seguente (supponiamo che il modello includa una costante):

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_{N-1} = 0 \quad \left( \sum_{i=1}^N \alpha_i = 0 \Rightarrow \alpha_N = 0 \right)$$

$H_1$  : non  $H_0$

$$\frac{(S_0 - S_1) / (N - 1)}{S_1 / (NT - N - K)} \stackrel{H_0}{\sim} F_{(N-1), (NT-N-K)}$$

<sup>3</sup>Si noti che questa condizione esclude la possibilità che vi siano variabili ritardate tra le variabili esplicative:  $y_{i,t-1}$  è infatti chiaramente correlata con  $u_{i,t-1}$ .

$S_0$  : somma del quadrato dei residui ottenuti stimando con OLS il modello ristretto, ovvero il modello con effetti “comuni” ( $H_0$ )

$S_1$  : somma del quadrato dei residui ottenuti dal modello che include le dummies ( $H_1$ ), stimato con LSDV oppure, se  $N$  è elevato, dal modello stimato nello stadio 1 con OLS.

Si rigetta l’ipotesi nulla se il valore della statistica risulta superiore al valore critico della distribuzione  $F$  con i rilevanti gradi di libertà al livello di significatività scelto.

Infine, si noti che lo stimatore  $\widehat{\beta}_{FE}$  è anche detto stimatore *within* perché è identificato attraverso la variabilità “interna” a ogni individuo:  $\widehat{\beta}_{FE}$  spiega, infatti, la misura in cui  $y_{it}$  differisce dalla propria media temporale,  $\bar{y}_i$  (si veda la trasformazione adottata nel primo stadio). Ciò implica che è impossibile di stimare i coefficienti di variabili che non variano nel tempo. Ad esempio, se la variabile dipendente è il salario e tra le variabili esplicative vogliamo includere il grado di istruzione o gli anni di studio, questa variabile rimarrà costante nel tempo per lo stesso individuo (supponendo si tratti di un individuo adulto). Consideriamo, dunque, un modello più generale che comprende anche variabili di questo tipo:

$$y_{it} = \alpha_i + x'_{it}\beta + z'_i\delta + u_{it}$$

La trasformazione effettuata nel primo stadio elimina le variabili  $z_i$  dalla regressione, così come elimina le  $\alpha_i$ , rendendone così impossibile la stima dei coefficienti.

## 2.2 Effetti individuali casuali

$$\begin{aligned} y_{it} &= x'_{it}\beta + \varepsilon_{it} \quad i = 1, \dots, N; \quad t = 1, \dots, T \\ \varepsilon_{it} &= \alpha_i + u_{it} \\ u_{it} & i.i.d.N(0, \sigma_u^2) \quad (N.B. \text{ su } i \text{ e } t) \\ \alpha_i & i.i.d.N(0, \sigma_\alpha^2) \quad (N.B. \text{ su } i) \\ u_{it} \text{ e } \alpha_i & \text{ indipendenti} \end{aligned}$$

L’insieme di parametri da stimare è diverso dal caso precedente:  $(\beta, \sigma_\alpha^2, \sigma_u^2)$ . Vediamo innanzitutto come è fatta la matrice di varianza-covarianza dell’errore.

$$V(\varepsilon_{it}) = V(\alpha_i + u_{it}) = \sigma_\alpha^2 + \sigma_u^2$$

$$cov(\varepsilon_{it}, \varepsilon_{is}) = E[(\alpha_i + u_{it})(\alpha_i + u_{is})] = \sigma_\alpha^2 \quad \forall t, s \quad (\rightarrow \text{ correlazione seriale per ogni individuo})$$

$$cov(\varepsilon_{it}, \varepsilon_{js}) = E[(\alpha_i + u_{it})(\alpha_j + u_{js})] = 0 \quad \forall i \neq j, t, s \quad (\rightarrow \text{ assenza di correlazione tra individui})$$

Dunque:

$$\begin{aligned} E(\varepsilon_i \varepsilon_i') &= \begin{pmatrix} \sigma_\alpha^2 + \sigma_u^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_u^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_u^2 & \dots & \sigma_\alpha^2 \\ \dots & \dots & \dots & \dots & \dots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 + \sigma_u^2 \end{pmatrix} = \underset{(TxT)}{V} \\ E(\varepsilon \varepsilon') &= \begin{pmatrix} V & 0 & 0 & \dots & 0 \\ 0 & V & 0 & \dots & 0 \\ 0 & 0 & V & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & V \end{pmatrix} = \underset{(NTxNT)}{\Omega} \end{aligned}$$

La matrice di varianza-covarianza dell'errore è dunque diagonale a blocchi: si ha correlazione seriale tra gli errori dello stesso individuo in diversi istanti nel tempo (dovuta alla presenza di  $\alpha_i$ ) e varianza costante nel tempo e tra individui diversi.

Data la struttura della matrice di varianza-covarianza dell'errore, lo stimatore OLS:

1. è non distorto e consistente se:  $E(\varepsilon_{is}|x_{it}) = 0, \forall i, s, t,$   
 è solo consistente se:  $E(x_{it}\varepsilon_{is}) = 0, \forall i, s, t$
2. non è più BLUE
3.  $V(\hat{\beta}_{OLS}) = (X'X)^{-1} (X'\Omega X) (X'X)^{-1}$ , ovvero diversa dalla formula utilizzata dal software per stimare la varianza.

Lo stimatore efficiente è lo stimatore GLS, anche detto stimatore *random effects* (RE):

$$\hat{\beta}_{GLS} = (X'\Omega X)^{-1} X'\Omega^{-1}y$$

Lo stimatore RE si ottiene pre-moltiplicando il modello per  $\Omega^{-\frac{1}{2}}$  e trasformando il modello nel modo seguente:

$$\begin{aligned} \Omega^{-\frac{1}{2}}y &= \{y_{it} - (1 - \lambda)\bar{y}_i\} \\ \lambda &= \sqrt{\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}} \\ [y_{it} - (1 - \lambda)\bar{y}_i] &= [x_{it} - (1 - \lambda)\bar{x}_i]' \beta + [\varepsilon_{it} - (1 - \lambda)\bar{\varepsilon}_i] \end{aligned}$$

L'errore del modello trasformato è omoschedastico e non autocorrelato (provate a verificarlo)  $\Rightarrow$  il modello trasformato può essere stimato con OLS.

Come sempre, lo stimatore GLS così ottenuto è "ideale" perché  $\sigma_\alpha^2$  e  $\sigma_u^2$  non sono noti. Dobbiamo quindi ripiegare su uno stimatore "fattibile", FGLS, ottenuto sostituendo alle quantità non note delle loro stime consistenti. Abbiamo già trovato uno stimatore consistente per  $\sigma_u^2$ :

$$\hat{\sigma}_u^2 = s_{FE}^2 \xrightarrow{p} \sigma_u^2$$

Occorre ora trovare uno stimatore consistente per  $\beta$ , che ricaviamo stimando con OLS il modello *between*:

$$\bar{y}_i = \bar{x}_i' \beta + \bar{\varepsilon}_i \quad i = 1, \dots, N$$

Otteniamo così lo stimatore *between*,  $\hat{\beta}_B$ , così detto perché identificato utilizzando esclusivamente la variabilità tra individui, per ognuno dei quali viene utilizzata solo la media temporale di ciascuna variabile. Questo stimatore è consistente (per  $N \rightarrow \infty$ ) se  $E(\bar{x}_i \alpha_i) = 0$  e  $E(\bar{x}_i \bar{u}_i) = 0$  ovvero se le variabili esplicative non correlate con tutte le  $u_{it}$  e con l'effetto individuale  $\alpha_i$ .

Si noti che l'errore nella regressione *between* è:

$$\bar{\varepsilon}_i = \alpha_i + \bar{u}_i$$

Ne consegue che<sup>4</sup>:

$$E(\bar{\varepsilon}_i) = 0$$

$$V(\bar{\varepsilon}_i) = V(\alpha_i + \bar{u}_i) = \sigma_\alpha^2 + \frac{\sigma_u^2}{T}$$

$$p \lim s_B^2 = \sigma_\alpha^2 + \frac{\sigma_u^2}{T} \Rightarrow P \lim \left( s_B^2 - \frac{\hat{\sigma}_u^2}{T} \right) = \sigma_\alpha^2$$

dove  $\hat{\sigma}_u^2$  è lo stimatore ottenuto dal modello FE, consistente per  $\sigma_u^2$ . Lo stimatore consistente per  $\sigma_\alpha^2$  da utilizzare nella trasformazione del modello ( $\leftrightarrow$  nella stima GLS) è:

$$\hat{\sigma}_\alpha^2 = s_B^2 - \frac{\hat{\sigma}_u^2}{T}$$

Riassumendo, lo stimatore *random effects* si ottiene stimando con OLS il modello:

$$\tilde{y}_{it} = \tilde{x}'_{it}\beta + \tilde{\varepsilon}_{it}$$

dove:

- $\tilde{y}_{it} = \left\{ y_{it} - \left( 1 - \sqrt{\frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + T\hat{\sigma}_\alpha^2}} \right) \bar{y}_i \right\}$  ( $\tilde{x}_{it}$  ed  $\tilde{\varepsilon}_{it}$  sono definite in modo analogo)
- $\hat{\sigma}_u^2 = s_{FE}^2$  e  $\hat{\sigma}_\alpha^2 = s_B^2 - \frac{\hat{\sigma}_u^2}{T}$

Si noti che, in alcuni casi,  $\hat{\sigma}_\alpha^2$  potrebbe risultare in una stima negativa (che, evidentemente, non ha alcun senso). Ciò accade quando  $\sigma_\alpha^2$  è prossimo a zero, cioè quando l'eterogeneità non è importante. In questi casi si considera dunque  $\sigma_\alpha^2 = 0$  e lo stimatore RE coincide con un semplice OLS sul modello originario ( $\sigma_\alpha^2 = 0 \Rightarrow \lambda = 1 \Rightarrow \tilde{y}_{it} = y_{it}$ ). Infatti, se  $\sigma_\alpha^2 = 0$  non c'è correlazione tra gli errori per lo stesso individuo e la matrice di varianza-covarianza del modello lineare originario è  $E(\varepsilon\varepsilon') = \sigma_u^2 I$ .

Si noti anche che  $T \rightarrow \infty \Rightarrow \lambda \rightarrow 0$  e quindi:

$$\tilde{y}_{it} = \{y_{it} - \bar{y}_i\}$$

In altri termini, gli stimatori FE e RE sono equivalenti per T molto grande. Più in generale, esiste una stretta relazione tra lo stimatore RE, lo stimatore FE ed anche lo stimatore between. Si può, infatti, dimostrare che:

$$\hat{\beta}_{RE} = \Delta \hat{\beta}_B + (I_K - \Delta) \hat{\beta}_{FE}$$

ovvero lo stimatore RE è una media ponderata dello stimatore *within* e dello stimatore *between*, dove i pesi sono legati alla varianza relativa dei due stimatori: quanto più uno stimatore è accurato ( $\leftrightarrow$  minore è la sua varianza), maggiore è il peso che gli viene assegnato. Lo stimatore RE è la combinazione ottima dello stimatore *within* e dello stimatore *between* ed è quindi più efficiente di entrambi<sup>5</sup>.

<sup>4</sup>L'errore è anche serialmente correlato per lo stesso individuo, ciononostante noi siamo interessati ad avere una stima consistente di  $\sigma_\alpha^2$  e l'OLS su questo modello, sebbene inefficiente, soddisfa questa richiesta.

<sup>5</sup>Anche lo stimatore OLS è una combinazione lineare degli stimatori *within* e *between*, ma non è quella efficiente.

Oltre ad essere efficiente, lo stimatore RE è non distorto se:

$$E(u_{is}|x_{it}) = 0 \text{ e } E(\alpha_i|x_{it}) = 0, \forall i, s, t$$

(cioè le variabili esplicative sono indipendenti da ogni  $u_{is}$  e ogni  $\alpha_i \leftrightarrow$  sono esogene). Perché  $\hat{\beta}_{RE}$  sia consistente (per  $N$  o  $T$  o entrambi tendenti ad infinito) è sufficiente che :

$$E(x_{it}u_{is}) = 0 \text{ e } E(x_{it}\alpha_i) = 0, \forall i, s, t$$

Infine, se valgono alcune condizioni di regolarità, si può dimostrare che  $\hat{\beta}_{RE}$  ha una distribuzione asintotica normale.

### 2.3 Effetti individuali fissi o casuali?

Gli stimatori FE e RE utilizzati su campioni con  $T$  piccolo e  $N$  grande possono dare origine a stime anche molto diverse. E' dunque opportuno chiedersi quale dei due stimatori sia più appropriato: l'effetto individuale  $\alpha_i$  deve essere considerato fisso o casuale? Questa è una domanda a cui non è facile dare risposta, perché solitamente la questione da affrontare non riguarda tanto la vera natura dell'effetto individuale, quanto piuttosto il tipo di dati di cui si dispone.

Iniziamo con l'osservare che i modelli stimati con FE e RE spiegano la variabile dipendente in modo "diverso": poiché nella stima FE gli effetti individuali sono considerati fissi, essi sono de facto inclusi tra le variabili esplicative in qualità di "costanti individuali", al contrario nella stima RE gli effetti individuali sono una componente dell'errore. Dunque, assumendo che vi sia indipendenza tra le variabili esplicative e tutti i termini di errore sia nel modello FE che nel modello RE:

$$\begin{aligned} FE & : y_{it} = \alpha_i + x'_{it}\beta + u_{it} \Leftrightarrow E(y_{it}|x_{it}, \alpha_i) = \alpha_i + x'_{it}\beta \\ RE & : y_{it} = x'_{it}\beta + \varepsilon_{it} \Leftrightarrow E(y_{it}|x_{it}) = x'_{it}\beta \end{aligned}$$

L'approccio FE è condizionale ai valori degli  $\alpha_i$ . Per questo motivo, risulta appropriato quando gli individui nel campione sono individui "particolari" e non possono essere pensati come estrazioni casuali da una popolazione. Ciò accade, ad esempio, quando  $i$  indica stati o regioni (come spesso accade nei panel macroeconomici), grandi imprese (es. multinazionali), settori industriali. In tutti questi casi, le inferenze che possiamo trarre sono necessariamente condizionali (e relative) agli individui inclusi nel campione. Diverso è il caso in cui gli individui nel campione possono essere pensati come estrazioni casuali da una popolazione: qui le caratteristiche individuali diventano una componente della variabilità della popolazione e le inferenze da un approccio RE sono quindi relative alla popolazione stessa.

In breve, una prima ragione per cui lo stimatore FE può essere preferito allo stimatore RE risiede nell'interesse verso gli  $\alpha_i$ : questo tipicamente esiste (ed ha senso) se gli individui nel campione sono in numero relativamente ridotto ( $N$  non è 10000!) ed hanno natura specifica, cosicché la loro identificazione è importante.

Ciononostante, esistono situazioni in cui l'approccio FE risulta preferibile anche se il numero di individui nel campione è relativamente elevato e siamo interessati ad inferenze sulla popolazione. Ciò accade quando  $\alpha_i$  e  $x_{it}$  sono correlati. Ad esempio, consideriamo nuovamente il panel contenente dati sull'output e sugli inputs impiegati da 576 imprese del settore manifatturiero per il periodo 1985-1994: se gli  $\alpha_i$  riassumono informazioni su caratteristiche individuali delle imprese quali le abilità manageriali, la cultura

e la struttura organizzativa, ecc., è ragionevole supporre che essi risultino correlati con le variabili di input. In questi casi, l'approccio RE fornisce stimatori inconsistenti mentre lo stimatore FE, che è ottenuto eliminando gli  $\alpha_i$  dal modello, continua ad essere consistente.

Supponiamo dunque che  $\alpha_i$  e  $x_{it}$  siano correlati:  $E(x_{it}\alpha_i) \neq 0$ . L'inconsistenza dello stimatore RE risulta evidente se scriviamo l'errore del modello trasformato:

$$RE : \varepsilon_{it} - (1 - \lambda)\bar{\varepsilon}_i = u_{it} - (1 - \lambda)\bar{u}_i + \lambda\alpha_i$$

L'errore contiene ancora  $\alpha_i$  e quindi risulta correlato con le variabili esplicative. Lo stimatore  $\hat{\beta}_{FE}$ , al contrario, è ottenuto stimando con OLS il modello in deviazione dalle medie individuali dove il termine di errore è:

$$FE : \varepsilon_{it} - \bar{\varepsilon}_i = u_{it} - \bar{u}_i$$

E' evidente che la consistenza di  $\hat{\beta}_{FE}$  non dipende in alcun modo dalla relazione tra gli effetti individuali e le variabili esplicative perché gli  $\alpha_i$  non sono inclusi nell'errore del modello stimato.

Per i motivi sopra spiegati, un test dell'ipotesi di non correlazione tra le variabili esplicative e gli effetti individuali è anche un test sull'affidabilità dello stimatore RE. Il test impiegato a questo scopo è il ben noto test di Hausman.

### Test di Hausman

L'idea generale del test di Hausman consiste nel confrontare due stimatori uno dei quali è consistente sia sotto l'ipotesi nulla di non correlazione che sotto l'ipotesi alternativa, mentre l'altro è consistente (ed efficiente) solo sotto l'ipotesi nulla e inconsistente sotto l'ipotesi alternativa. Le due ipotesi sono:

$$H_0 : E(x_{it}\alpha_i) = 0 \text{ vs } E(x_{it}\alpha_i) \neq 0$$

Sotto l'ipotesi nulla:

1.  $\hat{\beta}_{FE}$  è consistente:  $P \lim \hat{\beta}_{FE} = \beta$
2.  $\hat{\beta}_{RE}$  è consistente:  $P \lim \hat{\beta}_{RE} = \beta$
3.  $\hat{\beta}_{RE}$  è efficiente

Da 1 e 2 ricaviamo che:  $p \lim(\hat{\beta}_{FE} - \hat{\beta}_{RE}) = 0$ . Dunque il test può essere basato sulla differenza  $(\hat{\beta}_{FE} - \hat{\beta}_{RE})$ . Se questa risulta significativamente diversa da zero l'ipotesi nulla deve essere rigettata in favore di  $H_1$  ( $\leftrightarrow$  FE è consistente, mentre RE è inconsistente). Per comprendere com'è costruita e distribuita la statistica del test di Hausman occorre ricordare che:

a) Sotto  $H_0$  entrambi gli stimatori sono distribuiti asintoticamente secondo una normale  $\Rightarrow$  anche la loro differenza lo è:

$$\left. \begin{array}{l} \hat{\beta}_{FE} \overset{a}{\sim} N(\beta, V(\hat{\beta}_{FE})) \\ \hat{\beta}_{RE} \overset{a}{\sim} N(\beta, V(\hat{\beta}_{RE})) \end{array} \right\} \Rightarrow (\hat{\beta}_{FE} - \hat{\beta}_{RE}) \overset{a}{\sim} N(\beta, V(\hat{\beta}_{FE} - \hat{\beta}_{RE}))$$

b) Poiché sotto l'ipotesi nulla lo stimatore RE è efficiente, è possibile dimostrare che:

$$V(\hat{\beta}_{FE} - \hat{\beta}_{RE}) = V(\hat{\beta}_{FE}) - V(\hat{\beta}_{RE})$$

La statistica del test di Hausman è la seguente forma quadratica:

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})' \left[ \hat{V}(\hat{\beta}_{FE}) - \hat{V}(\hat{\beta}_{RE}) \right]^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE}) \sim \chi_K^2$$

dove  $\hat{V}$  indica la stima della matrice di varianza-covarianza vera di ciascun stimatore e  $K$  è il numero di elementi inclusi in  $\beta$ .

Alcune note sul test di Hausman:

- La varianza stimata di  $(\hat{\beta}_{FE} - \hat{\beta}_{RE})$ , ovvero  $\hat{V}(\hat{\beta}_{FE}) - \hat{V}(\hat{\beta}_{RE})$ , può non risultare definita positiva e quindi non invertibile: in questo caso, il test non può essere fatto. In alternativa si può fare il test su un sottoinsieme dei parametri inclusi in  $\beta$ .
- Se il modello include delle variabili individuali che rimangono costanti nel tempo, poiché i coefficienti di queste variabili non sono stimati con l'approccio FE, il test di Hausman confronta solo i  $\beta$ .

## 2.4 Bontà della stima

Per valutare la bontà della stima nei modelli visti nelle sezioni precedenti si utilizza la definizione dell' $R^2$  come il quadrato del coefficiente di correlazione tra i valori effettivi e quelli "fittati". Questa definizione assicura che i valori dell' $R^2$  così ottenuti siano compresi nell'intervallo  $[0, 1]$  e corrisponde alla definizione standard dell' $R^2$  nel modello OLS (se è inclusa la costante).

Nei modelli con dati panel è possibile mostrare che la varianza complessiva nelle  $y_{it}$  può essere scritta come la somma della varianza "within" e della varianza "between":

$$\frac{1}{NT} \sum_{i,t} (y_{it} - \bar{y})^2 = \frac{1}{NT} \sum_{i,t} (y_{it} - \bar{y}_i)^2 + \frac{1}{NT} \sum_{i,t} (\bar{y}_i - \bar{y})^2$$

dove  $\bar{y} = \frac{1}{NT} \sum_{i,t} y_{it}$ .

E' dunque possibile costruire tre diverse misure di  $R^2$ , con riferimento ai valori "fittati" dalle regressioni between, within e OLS:

1. Between:  $\hat{y}_i = \bar{x}_i' \hat{\beta}_B$   $R_{BETWEEN}^2 = \text{corr}^2 \left\{ \bar{x}_i' \hat{\beta}_B, \bar{y}_i \right\}$
2. Within:  $\hat{y}_{it} - \hat{y}_i = (x_{it} - \bar{x}_i)' \hat{\beta}_{FE}$   $R_{WITHIN}^2 = \text{corr}^2 \left\{ (x_{it} - \bar{x}_i)' \hat{\beta}_{FE}, (y_{it} - \bar{y}_i) \right\}$
3. OLS:  $\hat{y}_{it} = x_{it}' \hat{\beta}$   $R_{OVERALL}^2 = \text{corr}^2 \left\{ x_{it}' \hat{\beta}_{OLS}, y_{it} \right\}$

In realtà, i tre diversi  $R^2$  possono essere calcolati per un qualsiasi stimatore  $\hat{\beta}$ , utilizzando le formule precedenti in cui  $\hat{\beta}$  compare al posto dei vari  $\hat{\beta}_B, \hat{\beta}_{FE}, \hat{\beta}_{OLS}$  e i valori "fittati" sono:

$$\begin{aligned} \hat{y}_{it} &= x_{it}' \hat{\beta} \\ \hat{y}_i &= \frac{1}{T} \sum_t \hat{y}_{it} \\ \hat{y} &= \frac{1}{NT} \sum_{i,t} \hat{y}_{it} \end{aligned}$$

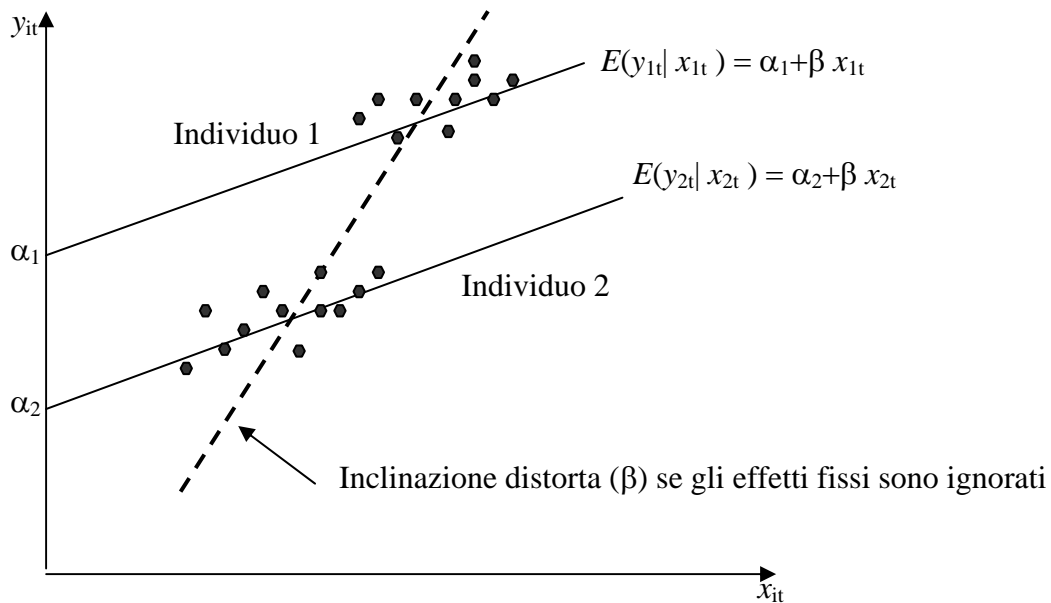


Figura 1. Distorsione dello stimatore OLS se  $\alpha_1 \neq \alpha_2$